Classifying and Extracting Data from Facebook Posts for Online Persona Identification

Hazel Anne Brosas
De La Salle University
Manila, Philippines
hazel_brosas@dlsu.edu.ph

Eugene Lim
De La Salle University
Manila, Philippines
eugene_lim@dlsu.edu.ph

Danica Sevilla
De La Salle University
Manila, Philippines
danica_sevilla@dlsu.edu.ph

Denise Anne Silva
De La Salle University
Manila, Philippines
denise_silva@dlsu.edu.ph

Ethel Ong
De La Salle University
Manila, Philippines
ethel.ong@delasalle.ph

Abstract

Large amount of user-generated data are posted online in social media platforms, including user preferences, dining and leisure activities, events, news and personal blogs. This resulted in varying efforts to process social media data using NLP and ML algorithms for topic classification, sentiment analysis and detection, and events classification. Such information are problematic to process, as they tend to be short, informal, inconsistent, and are highly contextualized. A series of tasks is involved from collecting, pre-processing, classification and extraction before social media data can be used. In this study, we built a multi-class classifier model to process Facebook posts in order to identify a user's online persona based on his/her preferences. Information extraction is then applied to find relevant data from the classified posts that can be used to generate a description of the user's online persona. The classifier currently achieves an accuracy of 76.02% and an F1 score of 73.10% using 10-fold cross validation from a dataset containing 16,682 posts.

1 Introduction

Facebook is the primary online social networking platform with an average of 1.45 billion daily active users worldwide. It has catered to the communication and socialization of its users to achieve varying purposes, including personal, academics (Kramer, 2015; Prescott et al., 2015), and business transactions (Hutchings, 2012; Culnan et al., 2010). By building a public profile, users combine text, images and video to share numerous kinds of content ranging from personal stories and activities to events, news, memes, blog posts, and business-oriented posts, to reach a large audience base.

People's online activities, reflected through their participation, motivation and practices, can be used to construct their online persona (Queensberry, 2015). An online persona is defined to be the social identity that an Internet user establishes in online communities and websites (Carminati et al., 2013). The study of Zhao, Grasmuck and Jason Martin (2008) revealed three (3) modes of online identity construction. These are visual with the use of display pictures, photos, and wall posts; enumerating hobbies, interests, and favorites; and describing oneself through narratives. While the study of Soraj (2011) argues that the persona people exhibit in social media platforms may not necessarily reflect their true identity, it is still worth exploring how automatic Facebook data classification according to user preferences can be used in online persona identification. This is essential in various fields, e.g. business, literature, and computing technology, wherein an understanding of the characteristics of users and/or consumers can aid in personalizing services and presentation of relevant information.

Several research efforts have combined NLP and ML algorithms to process the large amount of user-generated data from this platform, specifically with the motive of finding trends or patterns about the user's shared participation and practices. Using Facebook data, mainly posts, several studies have investigated on topic classification (Benkhelifa and Laallam, 2016), sentiment analysis or detection (Setty et al., 2014), and life events classification (Kinsella et al., 2011; Hade et al., 2017).

In online persona identification, An et al. (2016) utilized the URL links in Facebook posts to construct personas aimed primarily at product and content marketing. Because of the data privacy policies of Facebook, several factors that contribute to persona identification, namely gender, hobbies and liked pages, were not utilized. Instead, they resorted to using data retrieved from AJ+, a tool that allows viewing of digital content in SNS, smart phones, and other smart gadgets. Persona preferences were represented as clusters obtained by performing K-means++ clustering and listing of top 100 domains, yielding seven clusters of common interests. Another study by Tsai et al. (2015) performed concept semantic analysis on the user's Facebook posts, likes and shares, which are referred to as his/her activities and interests. Past and current actions were observed through the use of social behavioral patterns.

Working with user-generated data such as those found in Facebook poses some challenges, as they tend to be noisy (Petz et al., 2013; Dey and Haque, 2009; Abbasi et al., 2008) in nature. That is, texts are expressed in informal language and incomplete sentences. They may also contain incorrect grammar, misspelled words, emoticons, abbreviations and unnecessary capitalizations.

In this paper, we first present the ML techniques we used to build our user preference classifier models. We also present an analysis on the performance of the classifier models given the known issues in working with social media data. We then discuss the information extraction techniques we applied on the classified posts in order to extract data that may be relevant in describing a user's online persona. We end our paper with a brief discussion on using the extracted data as a knowledge source to generate a user's persona.

2 Building the Classifier Models

Several supervised ML algorithms were used in building the user preferences classifier models. The model with the highest F1 score is then used for identifying a user's online persona.

In the absence of scientifically validated persona labels, a set of user preference labels was first established by manually inspecting and annotating a small dataset comprising of 8,660 instances of text-based posts, liked pages and events. The resulting top five (5) persona labels is shown in Table 1.

Label	Description
The Fangirl/Fanboy	admires and stans celebrities or public figure
The Foodie	loves food and food-related activities like eating, cooking, etc.
The Gamer	plays games and interacts through it
The Melancholic	post revolves on hating on people, swearing; hates their life and everything in it; somber tone of status update
The Sports Fanatic	passionate for sport games and athletes

Table 1: Top Five (5) Persona Labels

The training dataset, containing 16,682 instances, was formed by combining the initial dataset used in establishing the user preference labels and collecting additional Facebook data. Both datasets were gathered with consent from individual Facebook users of 18 years of age and above. It is comprised of text-based posts, liked pages and events, which we refer to as Facebook posts in general, and written in mixed Filipino and English languages.

The text-based posts include the caption, and may optionally contain the name of the page or caption of the shared content. The liked pages and events include the name of the page/event and description of the page/event. The description is necessary because the name and category of the page/event cannot solely determine the type of user preference the page/event corresponds to.

Pre-processing was performed to clean the data,

but no normalization was performed. Because social media language is constantly changing, the available normalization tools for English and Filipino languages may not fully adapt to the dynamic content of posts. For example, it was observed that the use of *LOL* differs across user communities, wherein one context may imply *"laugh out loud"* while another context may mean *"League of Legends"*. Cleaning involves the removal of emojis, languages that use non-Latin alphabets, URLs, punctuations and digits, extra white spaces, and English and Filipino stop words. All texts were converted to lowercase to ensure that non-redundant features will be produced in feature extraction.

To further understand the textual content of each label in the dataset, the top frequently occurring words reflective of each persona label were generated. These are shown in Table 2. Based on the results, we have observed that the publicly available dictionaries of English and Filipino stop words are still insufficient to cover certain cases, such as "*shared*", "*official*", and "*page*" and "*mo*" (you), "*naman*" and "*yung*" (the), to name a few. These missing stop words that are particular to Facebook data were detected with the use of the top frequently occurring words and were manually added.

2.1 Multi-Class Classification

Features were extracted from the pre-processed texts together with the labels using word n-grams of 1, 2, 3, ranges of 1-2, ranges 1-3, and term frequency-inverse document frequency (tf-idf) weight calculation that includes a stemmer analyzer. Tf-idf can be computed by (Larson, 2010):

$$tf = \frac{\text{Number of times term } \mathbf{t} \text{ appear in a document}}{\text{Total number of terms in a document}}$$

$$idf = \log(\frac{\text{Total number of documents}}{\text{Number of documents with term } \mathbf{t} \text{ in it}})$$

$$tf - idf = tf \times idf$$

Two factors were considered in selecting n-grams as features. First, Facebook has strict data privacy policy that limits the available data to be extracted to mainly textual content. Secondly, our study also aims to determine the effectiveness of words in social media text as indicators of user preferences and online persona.

Since we are dealing with a multi-class classification problem, classifiers that apply One-Vs-The-Rest (OvR) strategy which is a fair default choice for multi-class classification rather than One-Vs-One (OvO) strategy (Pedregosa et al., 2011), were trained on the dataset and features. These are the Gradient Boosting Classifier (GBC), Linear SVC (LSVC), Logistic Regression (LR), Logistic Regression CV (LRCV), Stochastic Gradient Descent Classifier (SGDC), Perceptron (P), and Passive Aggressive Classifier (PAC). Among the evaluation methods, K-Folds Cross Validation (k=10) was performed as it better estimates the out-of-sample data (new and unseen).

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
GBC	72.76%	72.70%	72.76%	66.99%
LSVC	74.80%	73.71%	74.80%	72.22%
LR	72.23%	73.31%	72.23%	65.59%
LRCV	74.90%	73.65%	74.90%	70.85%
SGDC	74.87%	74.39%	74.87%	70.42%
P	71.03%	69.95%	71.03%	69.83%
PAC	72.97%	71.93%	72.97%	71.59%

Table 3: Performance of Multi-class Classifiers with Unigram Features

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
GBC	72.77%	72.72%	72.77%	67.00%
LSVC	74.70%	74.18%	74.70%	70.51%
LR	70.11%	73.55%	70.11%	61.13%
LRCV	74.49%	74.11%	74.49%	69.96%
SGDC	73.57%	74.71%	73.57%	67.61%
P	73.79%	72.22%	73.79%	71.93%
PAC	**75.08%**	**73.81%**	**75.08%**	**72.28%**

Table 4: Performance of Multi-class Classifiers with 1-2 N-grams Features

The models performed best with unigrams and 1-2 n-grams as features with accuracies ranging from 70% - 76% and F1 scores of 61% - 73% as shown in Tables 3 and 4, respectively. The low-performing models are those with bigrams and trigrams as fea-

#	Others	The Fangirl/ boy	The Foodie	The Gamer	The Melancholic	The Sports Fanatic
1	philippines	series	food	game	bored	team
2	manila	film	sm	games	dont	football
3	world	music	city	play	naman	sports
4	people	love	philippines	world	hate	nba
5	city	world	pizza	league	public	basketball
6	love	beat	mall	team	time	world
7	university	philippines	best	gaming	know	club
8	dlsu	album	starbucks	players	d	league
9	best	fan	chocolate	join	love	game
10	students	fans	center	dota	life	kobe

Table 2: Top 10 Words for each User Preferences Labels

tures having an accuracy range of 66% - 72% and F1 scores of 53% - 66%.

2.2 Selecting the Best Classifier Model

Out of all the 35 models trained, Passive Aggressive Classifier with 1-2 n-grams features performed the best, achieving a 75.08% accuracy and 72.28% F1-score. We also note that as the range of n-gram increases, the performance of the models decreases. The noisy nature of Facebook data may attribute to this performance degradation as 2 or 3 unique consecutive words appear less frequently in posts.

In choosing the best model, we have considered the accuracy paradox, which states that "high accuracy is not necessarily an indicator of high classifier performance" (Valverde, 2014). This phenomenon is highly associated to an imbalanced dataset which characterizes our case. Therefore, we decided to rely on the F1 score computed on macro-average (Van Asch, 2013) to utilize both precision and recall since F1 score provides the harmonic average between the two mentioned evaluation metrics:

$$B_{macro} = \frac{1}{q}\sum_{\lambda=1}^{q} B(tp_\eta \;\; fp_\eta \;\; tn_\eta \;\; fn)$$

Computing the F1 score based on macro-average instead of micro-average (traditional computation) gives equal weight to the classes. In this sense, we can obtain how effective the model is on small classes, rather than the large classes. Targeting the effectiveness of the model on small classes is necessary as the the top (5) user preferences labels are considered such.

2.3 Reducing the Features

As an additional experiment, classifier models were trained again, this time, with reduced features. Generally, feature reduction techniques in ML does not absolutely guarantee improvement in the classifier's performance, thus, validation is needed in this specific dataset and classification problem.

Since bigrams and trigrams showed significantly lower performances than other n-grams features, feature selection was only applied to unigrams, 1-2 n-grams, and 1-3 n-grams. Selecting the 10%, 20%, and 30% only of features using the ANOVA F-value and chi-squared test, hundred thousands of features were reduced. Based on the validation results, the classifier model actually performed better using only a small set of relevant features with Passive Aggressive Classifier still performing the best as highlighted in Table 5. With the 10% selected features using chi-squared in unigrams (3,810 in totality), it gained an accuracy of 76.02% and an F1-Score of 73.10%.

Model	Accuracy	Precision	Recall	F1 Score
GBC	72.15%	71.53%	72.15%	66.00%
LSVC	75.94%	75.33%	75.94%	72.54%
LR	71.71%	72.91%	71.71%	64.52%
LRCV	75.56%	74.89%	75.76%	72.20%
SGDC	73.36%	73.09%	73.36%	67.52%
P	72.48%	71.18%	72.48%	71.00%
PAC	**76.02%**	**74.73%**	**76.02%**	**73.10%**

Table 5: Performance of Multi-class Classifiers with Reduced 1 N-grams

According to Smialowski (2010), however, the performance of any model trained using supervised feature selection will experience overfitting unless evaluated against a new and unseen testing data.

3 Extracting Data from Posts

Relevant information for persona generation is extracted from a given set of classified posts by performing context understanding and then generating a set of data frames. Prior to information extraction, the classified posts are grouped according to their labels, since different approaches in extraction are applied per label. Data cleaning is also performed to reduce errors and inconsistencies in the data.

3.1 Data Cleaning

A Facebook post contains three main elements, namely event, shared post caption, and thoughts. *Event* comprises text that is automatically generated by Facebook to describe the user activity and is usually expressed as "<User>shared <Facebook page>'s post" and "<User>is at <Location>with <tagged people>." The *shared post caption* is the caption of the post that the user shared on his/her timeline. This element is included to provide context and clarify user sentiments that are sometimes vaguely stated or missing in the user's post. The *thoughts* element contains the user-generated content in a post's "*What's on your mind?*" field.

The *event* text is retained as is without undergoing any cleaning, to prevent introducing new errors, as illustrated in Table 6.

Original	USER shared Foodiverse's post
Cleaned	USER shared Universidad's post.

Table 6: Results of Cleaning the Event

Mild cleaning is applied for shared post caption, which involves removing hashtags, mentions and tagged people, and translating Tagalog words to English. There are three reasons behind this. First, some captions are long, consuming a considerable amount of processing time during cleaning. Secondly, it was also observed that when the shared post caption undergoes thorough cleaning, new errors would be introduced, as illustrated in Table 7. Lastly, the caption is only needed to support context understanding of the user's thoughts, as discussed in Section 3.2 Context Understanding.

Thorough cleaning, on the other hand, was applied to *user thoughts* and involves a series of steps. These are the removal of hashtags and mentions; extraction of tagged people; removal of HTML tags; standardization of words; removal of expressions like "Haha" and "Hehe", newlines, emoticons, special characters, URLs and email addresses; apostrophe lookup; slang lookup; spelling corrector; and translation from Tagalog to English. Only the *user thought* element of a post undergoes this extensive cleaning process because it is the primary source for information extraction, wherein the inconsistencies in the grammar structure and the presence of informal words would make it difficult for context understanding to determine the message that the user is trying to convey. Table 8 shows the results of applying thorough cleaning on a sample *user thought*.

Thorough Cleaning	oreo cookie cheesecake! this must be a dream? ? follow **Universidad** for a whole new food **password**
Mild Cleaning	OREO COOKIE CHEESECAKE! This must be a dream? ? Follow **Foodiverse** for a whole new food **world** !.

Table 7: Results of Thorough vs Mild Cleaning of Shared Post Caption

Original Post	di ko na pala kailangan lumayo para sa korean food na yan hahaha?.
Cleaned Post	I do not have to go away for that korean food anymore?.

Table 8: Results of Thorough Cleaning of a User Thought

Despite the reduction in errors and inconsistencies in the labelled dataset after the cleaning process, there are still some challenges that need to be addressed. The attached words commonly found in hashtags should be split in order to utilize ad-

32nd Pacific Asia Conference on Language, Information and Computation
Hong Kong, 1-3 December 2018

ditional information that these can provide (Giannoulakis and Tsapatsoulis, 2016). The identification of slang words requires a dictionary that is constantly updated to keep up with the dynamic changes evident in the way people use slang words in their posts (Pedersen, 2007). Users also tend to express their feelings using emojis (Hakami, 2017). Processing such elements in a post can further increase the understanding on user thoughts.

3.2 Context Understanding

To extract relevant data from a user's post that describes his/her activities, context understanding involves determining the subject of the post, and then identifying other supporting details such as description, activity and subject type. Table 9 enumerates the specific list of information that will be extracted depending on a post's user preference label. The subject to be identified from the post based on its label is shown in **boldface**.

The Foodie	Person, **Food**, Adjective, Verb, Organization, Type, Friends, Location, Sentiment, SentClass, Date, Time
The Fangirl /Fanboy	Person, **FanOf**, Adjective, Verb, Event, Type, Friends, Organization, Location, Sentiment, SentClass, Date, Time
The Sports Fanatic	Person, **Sport**, Team, FanOf, Achievement, Verb, Event, Type, Friends, Organization, Location, Sentiment, SentClass, Date, Time
The Gamer	Person, **Game**, Adjective, Verb, FanOf, Team, Event, Type, Friends, Organization, Location, Sentiment, SentClass, Date, Time

Table 9: List of Information to be Extracted

There are four possible situations in finding the subject, depending on the user's purpose for making the post, i.e., providing a statement, expressing a sentiment, narrating an activity, and describing the subject. A post is identified as providing a statement when the user is claiming something about the subject. Consider the following post:

"*also. oldcodex performs theme song for all 3 compilation films opening this year.*"

The user is sharing details about *oldcodex*, which is identified to be the subject of the post. In this instance, the supporting details to be extracted are the subject, organization, type, tagged friends, date and time.

A post is identified as expressing a sentiment when the user shares his/her thoughts and feelings towards the subject. Consider the following post:

"*I am a bit disappointed with Sherlock Holmes.*"

The user is expressing a negative sentiment (*SentClass*) towards the subject, which is *Sherlock Holmes*. In this situation, the supporting details to be extracted are the subject, organization, type, sentiment, sentiment class, tagged friends, date and time. IBM Watson's sentiment analysis feature is used to determine the sentiment of a given word.

A post is identified as narrating an activity when the user performs an activity to the subject. This follows the transitive verb sentence structure in which the doer, action and receiver have to be distinguished. Given that user-generated texts are sometimes incomplete, there are cases when the doer cannot be determined. In this case, the doer is assumed to be the user. An example of this is the post "*Craving for Isaw*" where the subject is "*Isaw*" and the doer is the user because it is not explicitly stated. In this instance, the supporting details to be extracted are the subject, organization, type, verb, event, location, achievement, tagged friends, date and time.

Finally, a post such as "*the lion king is so far the best movie for me.*" is identified as describing a subject, when the subject "*the lion king*" is being described by the user as "*the best*". The supporting details to be extracted in this type of posts are the subject, organization, type, adjective, tagged friends, date and time.

3.3 Data Frames

The details extracted from the posts are stored into a data frame consisting of the following: person or user's name, subject, subject type, adjective describing the subject, verb describing the action, and other details as enumerated in Table 9. A sample data frame for a post with *The Foodie* label and with *food* as the subject is shown in Listing 1.

```
{
      "Person": "Hazel",
      "Food": "korean spicy noodles",
      "Adjective": none,
      "Verb": "crave",
      "Organization": none,
      "Type": ["grains and pasta"],
      "Tagged_Friends": none,
      "Location": none,
      "Sentiment": none,
      "SentClass": none,
      "Date": "June 20, 2017",
      "Time": "14:36",
}
```

Listing 1: Data frame for a post labeled *The Foodie*.

A data frame is used to construct assertions for a given user preference label. An assertion, in the literary concept, is defined as a positive statement regarding a belief or a fact. It could be transformed into statements that provide descriptions to support a user's identified online persona based on his/her user preferences. Each user preference label has its associated set of assertion types – five assertion types for *The Fangirl/Fanboy*, six for *The Gamer*, three for *The Foodie*, and six for *The Sports Fanatic*.

Consider the assertion types for *The Foodie* label shown in Table 10. For each parameter in an assertion type, the system looks for a corresponding parameter in the data frame. The associated value of the matching parameter is then used to instantiate an assertion type. The process is repeated until all assertion types for a given user preference label has been instantiated. Using the sample data frame in Listing 1, the resulting instantiated assertion types are shown in Listing 2.

```
food_describe("Hazel", "korean spicy noodles",
    null, null, "grains and pasta")
food_activity("Hazel", "korean spicy noodles",
    "crave", null, null, null, "grains and
    pasta")
food_sentiment("Hazel", "korean spicy noodles",
     null, null, "grains and pasta")
```

Listing 2: Instantiated assertion types for a post labeled *The Foodie*.

From the resulting assertion instances, only those assertion(s) whose required parameters contain values will be selected. With this rule, the assertion type *food_describe* is not selected since it lacks a value for one of its required parameters, i.e., the *description* of the food.

assertion type	**description**	**parameters**
food_describe	used to state a description of the food	Person, Food, Description, Organization, Type
food_activity	used to state the activity about food that a person is doing at a certain location with his/her friends	Person, Food, Action, Organization, Location, Tagged_Friend, Type
food_sentiment	used to state the sentiments of the person about the food	Person, Food, Sentiment, SentClass, Type

Table 10: Assertion Types for *The Foodie*

The extracted information is highly affected by mislabels and problems in the Facebook data. There are instances when the extracted information is not related at all to the identified user preference label. Consider the sample post below that has been labeled as *The Fangirl/Fanboy*:

"*it took me this long to realize i can just put another fan in my room*"

The post is not at all connected to *The Fangirl/Fanboy* user preference but it has been labeled as one because of the word "*fan*". Since the post can be identified as narrating an activity, the extracted information can be used to instantiate a *fan_activity* assertion type.

Furthermore, the missing context in some posts made it difficult to determine the subject and its associated details from the posts. The problem is compounded by the presence of contents such as song lyrics, lines from a movie, and quotes from a book.

4 Generating the Online Persona

A persona description for a given user has three components, namely personal information, overview, and user preferences. The *overview* describes the basic information of the Facebook user.

The *user preferences* contains a narrative to justify each of the identified persona.

The assertions derived from the extracted data are represented as RDF and used as the knowledge resource for a grammar-based generator. The story generator selects relevant assertions to form the persona description. Each persona label has an associated set of grammar rules to generate the story text. An example persona description generated for a user labelled as *The Foodie* is shown in Listing 3.

```
    She enjoys eating different cuisines
specifically philippine cuisine, japanese
cuisine and italian cuisine. For philip-
pine cuisine, she likes to eat isaw. For
japanese cuisine, she likes to eat ramen.
For italian cuisine, she likes to eat
pizza. Besides that, she enjoys eating
grains and pasta food like korean
spicy noodles. She likes oreo cheesecake
for dessert. She went to food places like
Bubblicitea Cafe.
```

Listing 3: Sample persona description for a user labeled as *The Foodie*.

The generated persona-based life story underwent end-user evaluation using two main criteria, namely, language and content. The language criterion assesses the correctness of spelling, punctuation, capitalization and grammar. While results show the presence of misspelled words, grammar errors and irrelevant information, users generally find the story to be creative and interesting. The content criterion assesses the generated online persona and the context of the story. Results show that the story contained information sourced from the user's Facebook posts, but on occasion, can miss a few things about the user and can even include information that is not applicable to the selected persona. For example, the text "*Game of Thrones Rocks My World.*", which makes reference to an American TV series, should not have been included as part of the description for a user whose persona is *The Gamer*.

5 Conclusion and Future Work

A user's online persona is expressed in his/her preferences which can be extracted from his/her social media posts, likes, interests and activities. In this study, a multi-class classifier model was built in order to identify the online persona based on user preferences in Facebook posts. Through the use of word/s (n-grams) found in social media data as features, the study investigated the effectiveness of the words or textual content of social media text as the primary indicator of one's preferences. Once labelled, relevant information are then extracted from the posts to produce data frames that serve as the knowledge resource for the generation of story-based persona description.

Validation results showed that the use of all extracted features will not necessarily lead a classifier model to perform at its best. Experiments conducted using only 10% of the features, specifically in the unigrams, revealed that the Passive Aggressive Classifier achieved the highest performance among the different models that were trained, with an accuracy of 76.02% and F1-Score of 73.10%. There is an approximately 1% increase in comparison with the performance of the best-performing model in the non-reduced features. However, the best-performing model of the reduced features should be evaluated against an out-of sample test set in order to know its true performance.

Future works could focus on improving the preprocessing done on the dataset by applying normalization techniques, which can provide additional information to help clarify the context of the post. The performance of the classifier model can also be improved through exploration of different ratios or percentages of reduced features and parameter tuning. The performance of the best-performing model(s) could be further validated through experiments that consider the FB posts on a per user perspective, to reduce the impact of an unbalanced dataset caused by heavy posting of certain users.

The correctness of the information extraction module largely depends on the correct classification of the posts and the availability of clean data. Mis-labelled posts present difficulty in the identification of the subject of the post. This could lead to the extraction of incorrect details, which in turn affects the quality of the assertions that will subsequently be used for the generation of persona description. Further works should also look into identifying more subject situations, expanding the labels to cover more user preferences, and improving the data cleaning performed on social media texts.

References

Ahmed Abbasi, Hsinchun Chen, and Salem Arab. 2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems (TOIS)*, 26(3), 12. http://doi.acm.org/10.1145/1361684.1361685.

Francisco J. Valverde-Albacete and Carmen Pelez-Moreno. 2014. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. *PloS one*, 9(1), e84217. https://doi.org/10.1371/journal.pone.0084217.

Jisun An, Hoyoun Cho, Haewoon Kwak, Ziyaad Haasen Mohammed, and Bernard Jansen. 2016. Towards Automatic Persona Generation Using Social Media. *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 206–211, August 22-24 2016, Vienna, Austria.

Vincent Van Asch. 2013. Macro-and Micro-Averaged Evaluation Measures.

Randa Benkhelifa and Fatima Zohra Laallam. 2016. Facebook Posts Text Classification to Improve Information Filtering. *The 12th International Conference on Web Information Systems and Technologies (WEBIST), 202–207, April 23-25 2016, Rome, Italy.*

Barbara Carminati, Elena Ferrari, and Marco Viviani. 2013. Security and Trust in Online Social Networks. Morgan and Claypool Publishers.

Mary Culnan, Patrick McHugh, and Jesus Zubillaga. 2010. How large US companies can use Twitter and other Social Media to Gain Business Value. *MIS Quarterly Executive*, 2010(9).

Lipika Dey and Mirajul Haque. 2009. Opinion Mining from Noisy Text Data. *International Journal on Document Analysis and Recognition (IJDAR)*, 205–226.

Stamatios Giannoulakis and Nicolas Tsapatsoulis. 2016. Evaluating the Descriptive Power of Instagram Hashtags. *Journal of Innovation in Digital Ecosystems*, 3(2), 114–129. https://doi.org/10.1016/j.jides.2016.10.001.

Alden Luc Hade, Janica Mae Lam, Camille Alexis Saavedra, Robee Khyra Mae Te, and Ethel Ong. Extracting and Classifying Events from Social Media Posts. *14th National Natural Language Processing Research Symposium*, May 11-12 2018, Baguio City, Philippines.

Shatha Ali Hakami. 2017. The Importance of Understanding Emoji. University of Birmingham, Research Topics in HCI.

Chris Hutchings. 2012. Commercial Use of Facebook and Twitter–Risks and Rewards. *Computer Fraud and Security*, 2012(6), 19–20.

Sheila Kinsella, Alexandre Passant, and John Breslin. 2011. Topic Classification in Social Media Using Metadata from Hyperlinked Objects. *Proceedings of the 33rd European conference on Advances in information retrieval (ECIR'11), Paul Clough, Colum Foley, Cathal Gurrin, Hyowon Lee, and Gareth J. F. Jones (Eds.)*. Springer-Verlag, Berlin, Heidelberg, 201–206.

Rojas C. Kramer, Esquivel I. Gamez, and Garcia A. Santillan. 2015. Educational Use of Facebook in Higher Education Environments; Current Practices and Guidelines. *The 9th International Technology, Education and Development Conference*, 6042–6052, March 2-4 2015, Madrid, Spain.

Ray R. Larson. 2010. Introduction to Information Retrieval. https://doi.org/10.1002/asi.21234. *Journal of the American Society for Information Science and Technology*, 61 (4), April 2010, 852–853. DOI=http://dx.doi.org/10.1002/asi.v61:4.

Tim Pedersen. 2007. The Use of Slang in British English - A Study of the Slang used in Football Factory and Little Britain. The University of Kalmar.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011(12), 2825–2830.

Gerald Petz, Michal Karpowicz, Harald Frschu, Andreas Auinger, Vclav Sttesk, and Andreas Holzinger. 2013. Opinion Mining on the Web 2.0 - Characteristics of User Generated Content and Their Impacts. In Holzinger A., Pasi G. (eds), *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (HC1-KDD), LNCS 7947*, 35–46, Springer, Berlin, Heidelberg.

Julie Prescott, Matthew Stodart, Gordon Becket, and Sarah Wilson. 2015. The Experience of using Facebook as an Educational Tool. *Journal of Health and Social Care Education* Routledge. https://doi.org/10.11120/hsce.2013.00033.

Keith Queensberry. 2015. Harness the Power of Personas for Social Media Marketing. Retrieved from http://www.marketingprofs.com/articles/2015/28302/harness-the-power-of-personas-for-social-media-marketing.

Shankar Setty, Rajendra Jadi, Sabya Shaikh, Chandan Mattikalli, and Uma Mudenagudi. 2014. Classification of Facebook News Feeds and Sentiment Analysis. *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 18–23, New Delhi, India.

Pawel Smialowski, Dmitrij Frishman, and Stefan Kramer. 2010. Pitfalls of Supervised Feature Selection. *Bioinformatics*, 26(3), 440–443.

Hongladarom Soraj. 2011. Personal Identity and the Self in the Online and Offline World. *Minds and Machines*, 21(4), 533–548. https://doi.org/10.1007/s11023-011-9255-x.

Hung Cheng Tsai, Wen Han Liu, Ping Che Yang, Tsun Ku, and Wu Fan Chien. 2015. Social Persona Preference Analysis on Social Networks. *2015 International Conference on Connected Vehicles and Expo (ICCVE)*, 32–39, Shenzhen, China.

Shanyang Zhao, Sherri Grasmuck, and Jason Martin 2008 Identity Construction on Facebook: Digital Empowerment in Anchored Relationships. *Computers in Human Behavior*, 24(5), 1816-1836. https://doi.org/10.1016/j.chb.2008.02.012.